Journal of Informetrics 4 (2010) 320–330

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Conjugate partitions in informetrics: Lorenz curves, h-type indices, Ferrers graphs and Durfee squares in a discrete and continuous setting

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium

ARTICLE INFO

Article history:
Received 25 November 2009
Received in revised form 27 January 2010
Accepted 30 January 2010

Keywords:
Conjugate partition
Informetrics
Lorenz curve
h-index
g-index
R-index
Ferrers graph
Durfee square

ABSTRACT

The well-known discrete theory of conjugate partitions, Ferrers graphs and Durfee squares is interpreted in informetrics. It is shown that partitions and their conjugates have the same h-index, a fact that is not true for the g- and R-index. A modification of Ferrers graph is presented, yielding the g-index.

We then present a formula for the Lorenz curve of the conjugate partition in function of the Lorenz curve of the original partition in the discrete setting.

Ferrers graphs, Durfee squares and conjugate partitions are then defined in the continuous setting where variables range over intervals. Conjugate partitions are nothing else than the inverses of rank-frequency functions in informetrics. Also here they have the same h-index and we can again give a formula for the Lorenz curve of the conjugate partition in function of the Lorenz curve of the original partition. Calculatory examples are given where these Lorenz curves are equal and where one Lorenz curve dominates the other one. We also prove that the Lorenz curve of a partition and the one of its conjugate can intersect on the open interval]0, 1[.

1. Introduction

The attention of this author was drawn by the paper Anderson et al. (2008) where, for partitions (to be explained below), Ferrers graphs and Durfee squares are used to define a variant of the h-index.

For the sake of completeness we will define these simple concepts here (see also Andrews (1998)). A partition is, simply, a vector of finite dimension T with decreasing coordinates which are elements of $\mathbb{N}$, the positive natural numbers (excluding 0). We will denote such a vector by $C = (c_1, c_2, \ldots, c_T)$ where $c_1 \geq c_2 \geq \ldots \geq c_T \geq 1$. The name partition comes from the fact that the natural number $\sum_{i=1}^{T} c_i$ is partitioned by C, whereby each coordinate $c_1, \ldots, c_T$ indicates the relative size of the coordinate $i = 1, \ldots, T$.

Informetrically one can interpret C as a description of the number of items c_i in the source i, $i = 1, \ldots, T =$ *the total number* of sources. In this interpretation,

$$A = \sum_{i=1}^{T} c_i \tag{1}$$

E-mail address: leo.egghe@uhasselt.be.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2010.01.006

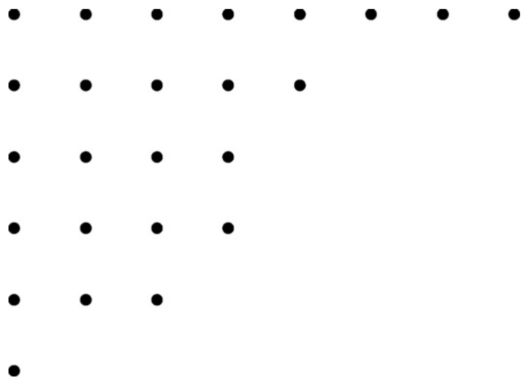

Fig. 1. Ferrers graph of $C = (8, 5, 4, 4, 3, 1)$.

is the total number of items in the system (also called an information production process (IPP) in which source i produces (or has) c_i items, $i = 1, \ldots, T$. In this interpretation, the vector C is nothing else than the rank-frequency function of this IPP. For more on IPPs we refer the reader to Egghe (2005) where many examples are given by interpreting C as an author or a journal and where the T articles of this author or journal are ranked in decreasing order of the number of citations c_i that these articles have received (e.g. until now), $i = 1, \ldots, T$.

In the theory of partitions, there is a handy graphical representation of a partition as a so-called Ferrers graph. Taking $C = (8, 5, 4, 4, 3, 1)$ as an example, the Ferrers graph of C is depicted as in Fig. 1.

The largest, fully filled, square in this graph (starting from the point on the first row and the first column) is called the Durfee square (after W.P. Durfee). The side of this Durfee square is nothing else than the h-index of this system (here $h = 4$). The h-index was introduced in Hirsch (2005) as the largest rank $r \in \{1, \ldots, T\}$ such that $c_r \geq r$ (hence $c_h \geq h$ and hence $c_i \geq h$ for all $i = 1, \ldots, h$ while $c_{h+1} < h + 1$). For more on the use and the (dis)advantages of the h-index we refer to the vast literature and the extensive review Egghe (2010). That the side of the Durfee square has size h was first remarked in Anderson et al. (2008).

In the theory of partitions one defines the notion of "conjugate" of a partition $C = (c_1, \ldots, c_T)$, cf. Andrews (1998), Definition 1.8. We define the conjugate of partition C as $C' = \left(c_1^{'}, \ldots, c_m^{'}\right)$ where for $j = 1, \ldots, m$, $c_j^{'}$ equals the number of coordinates in C that are larger than or equal to j. In the example above: $C = (8, 5, 4, 4, 3, 1)$ we see that its conjugate is $C' = (6, 5, 5, 4, 2, 1, 1, 1)$. It is easily seen (cf. Andrews (1998)) that C' is obtained from C by mirroring the Ferrers graph of C over the main diagonal (i.e. the line connecting the points for which the column number equals the row number)–see Fig. 2.

Note that dim C' (the dimension of C') does not always equal dim C. Only if c_1 = dim C we have that dim $C' = c_1$ and hence dim C = dim C'. Example: $C = (6, 5, 5, 4, 2, 1)$. Now $C' = (6, 5, 4, 4, 3, 1)$.

Note also that

$$\sum_{i=1}^{T} c_i = \sum_{j=1}^{m} c_j^{'} \tag{2}$$

In the informetric terminology of sources and items in IPPs we can define the conjugate of the rank-frequency function C as the function yielding for every $j = 1, \ldots, m$ the number of sources with $\geq j$ items

Ferrers graphs yield a trivial proof of the following important proposition

Proposition 1. *The h-index of a partition equals the h-index of its conjugate.*

Proof. In mirroring Ferrers graph of C over the main diagonal, the size of the Durfee square is not changed, hence it has the same side-length (hence h-index) as the one of C. Consequently $h_C = h_{C'}$.

One can indeed verify in the above example that $h_C = h_{C'} = 4$. Note that Ferrers graphs are handy in finding the conjugate C' of a partition C as well as in the proof of Proposition 1.

The above proposition is false for the g-index and the R-index as we will show below but first a short introduction to the g-index and the R-index. The h-index has the disadvantage of not counting the actual number of citations to papers in the h-core (the h-core is defined as the set of the first h papers; the term is only a convenient definition but does not always constitute a real core set of papers for the author or the journal). Indeed, once a paper is in the h-core, it does not matter how many citations ($\geq h$) this paper actually receives. This was remarked in Egghe (2006) where e.g. the example was given of equal h-indices for E. Garfield and F. Narin (early 2006), namely $h = 27$ while Garfield had 14 papers with more than 100 received citations and Narin had only 1 paper with more than 100 received citations (namely 112). Furthermore, Garfield had 1 paper with 625 received citations.

In order to improve the h-index for this (almost) insensibility to the actual number of received citations, we defined in Egghe (2006) the g-index as follows: in the same ranking as in $C = (c_1, \ldots, c_T)$, $r = g$ is the highest ranking such that the first